DeepVision-103K: A Visually Diverse, Broad-Coverage, and Verifiable Mathematical Dataset for Multimodal Reasoning

Anonymous ACL submission

Abstract

Reinforcement Learning with Verifiable Rewards (RLVR) has advanced the reasoning of vision language models (VLMs), yet remains bottlenecked by training data: synthetic corpora lack real-world scenarios, human-annotated sets are hard to scale, and recombined sources offer little distinct signal. To address these limitations, we introduce **DeepVision-103K**, a verifiable multimodal dataset for RLVR. Curated from real K12 sources through validity filtering, difficulty calibration, and correctness verification, it contains 103K visually diverse QA pairs spanning multimodal mathematics and visual logic. Models trained on DeepVision-103K achieve strong performance on both mathematical and general multimodal benchmarks across 4B–8B backbones. The 8B model matches the accuracy of its official thinking variant while using only ∼63% of the response tokens. Mixing a 20K subset of DeepVision-103K with existing datasets brings further gains, confirming its distinct value. Case-level analysis reveals three enhanced capabilities in DeepVision-trained models: visual perception, visual reflection, and mathematical reasoning.

1 Introduction

Reinforcement learning with verifiable rewards (RLVR) has been widely applied to vision language models (VLMs), yielding great improvements in visual reflection and reasoning (Wang et al., 2025a; Xia et al., 2025; Yang et al., 2025a). Despite this promise, the effectiveness of RLVR is constrained by existing training data which fall short in several key aspects. (1) **Narrow visual scope**: synthetic datasets (Lu et al., 2021; Qiao et al., 2025) curated with tools like GeoGebra focus on pure mathematical scenarios (e.g., geometric figures and function plots), lacking real-world scenarios; (2) **Poor scalability**: human-annotated K12 datasets (Meng et al., 2025; Liu et al., 2024) from authentic educational data covers broader visual content, but expert annotation caps their scale, leaving them short of RLVR's data needs.; (3)**No novel signal**: recombined datasets (Wang et al., 2025c; Zha et al., 2025; Peng et al., 2025; Yang et al., 2025b; Zhang et al., 2025) only filter or aggregate existing sources, providing no new training signal. These limitations hinder RLVR-trained models from acquiring robust and generalizable reasoning capabilities.

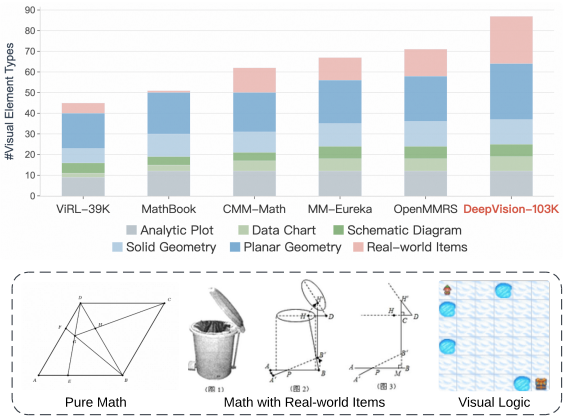


Figure 1: **Top**: Visual element coverage across RLVR datasets; **Bottom**: Examples of pure math, math with real-world items, and visual logic.

To this end, we propose **DeepVision-103K**, a large-scale multimodal mathematical dataset designed for RLVR, featuring:

- **Real-world Visual Contexts**: DeepVision-103K covers diverse pure mathematical scenarios and **math problems in real-life visual contexts** (Figure 1), requiring reasoning across varied visual settings.

- **Scalable Curation**: an automatic three-stage pipeline (validity filtering, pass-rate stratification, correctness verification) produces 103K verifiable QA pairs without expert annotation, delivering consistent gains across 4B–8B backbones (Figure 2 & Table 3).

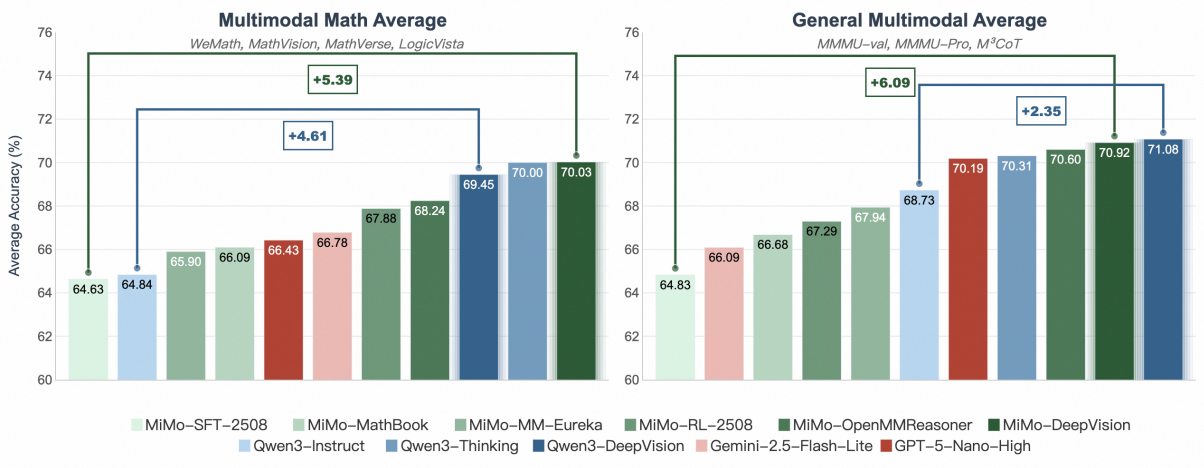


Figure 2: Performance on math and general benchmarks, we report averaged Pass@1 accuracy across benchmarks.

- **Novel Training Signals**: DeepVision-103K combines two domains—**a math subset** filtered from a 3.3M-sample raw K12 pool, and **a visual logic subset** (mazes, chess, tetris) for cross-domain coverage. Adding 20K DeepVision queries to existing RL datasets brings clear gains (Sec. 5.1), confirming its non-overlapping signal.

Models trained on DeepVision-103K show three advantages: (1) **Consistent improvements**: DeepVision training yields gains on both math/general benchmarks across 4B–8B backbones (+4.73/+5.66 on Qwen3-VL-4B, +4.61/+2.35 on Qwen3-VL-8B, +5.39/+6.09 on MiMo-VL-7B), matching or surpassing official thinking variants and closed-source baselines. (2) **Superiority over existing datasets**: Under the same RL settings, models trained on DeepVision-103K outperform those trained on three open-source RL datasets by +1.79 to +4.13 on math/general benchmarks. (3) **Reasoning efficiency**: On Qwen3-VL-8B serie, DeepVision model attains comparable accuracy to the official thinking variant (70.15% vs. 70.13%) while consuming only 63.5% tokens (Table 4). Our further analyses reveal several key findings: (1) **Distinct data value**: mixing a 20K subset of DeepVision-103K with existing open-source datasets yields further improvements, confirming that DeepVision provides non-overlapping training signal (Sec. 5.1); (2) **Math–logic synergy**: multimodal math and visual logic each contribute unique value to visual reasoning, and their benefits compound when combined (Sec. 5.2); (3) **Enhanced meta-behaviors**: case-level human annotation identifies enhanced visual perception, reflection, and mathematical reasoning in DeepVision models (Sec. 5.4).

2 Overview of DeepVision-103K

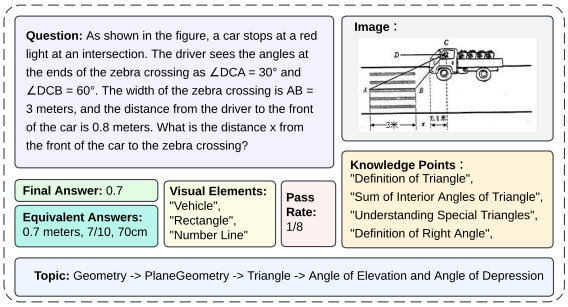


Figure 3: A data sample from DeepVision-103K.

Each sample in DeepVision-103K is annotated with seven fields supporting difficulty-aware filtering, topic-level analysis, and rule-based reward computation for RLVR.

Field	Description
Question & Image	The textual problem statement and its image.
Final Answer	A unique, verifiable answer that enables rule-based reward computation in RLVR.
Equivalent Answers	Alternative forms of the final answer, used to make rule-based matching more robust.
Pass Rate	Fraction of correct responses during model rollouts.
Topic	Hierarchical label of the mathematical branch the problem belongs to.
Knowledge Points	Mathematical concepts, theorems, or techniques required to solve the problem.
Visual Elements	Geometric or graphical objects in the image.

Table 1: Annotation fields and definitions.

2.1 Visual Diversity

To characterize the visual content of DeepVision-103K, we built a taxonomy based on Mo et al. (2018); Rosin (2008) and prompted GPT-5-mini to label the visual elements in each image with *categories* and *fine-grained types*. Prompts and details are provided in Appendix B. Under this taxonomy, DeepVision-103K covers six categories and contains more fine-grained element types than existing open-source datasets (Figure 1).

Table 2 summarizes the key elements in DeepVision-103K. Two aspects stand out: **cross-category combinations**, where a single image requires reasoning over multiple visual categories, and **real-world items in mathematical contexts**, which are rarely covered by existing datasets. Examples are provided in Figures 10 and 11.

Category	Key Visual Elements
`Planar Geometry`	Primitives (Angle, Triangle, Circle, Quadrilateral,Polygon), Relations (Parallelism, Tangency, Chords), Properties (Right Angles, Perpendicularity)
`Solid Geometry`	3D Primitives (Cube, Prism, Cylinder, Cone), Spatial Representations (Orthographic Views, Nets), Sections (Frustums, Hemispheres)
`Analytic Plot`	Coordinate Systems, Function Curves (Linear, General), Conic Sections (Parabola, Hyperbola), Scatter Points, Inequality Regions
`Data Chart`	Statistical Graphs (Bar, Histogram, Pie, Line), Structured Data (Tables, Stem-and-Leaf)
`Schematic Diagram`	Logical Structures (Flowcharts, Tree Diagrams), Physics/Sets (Force Diagrams, Circuits, Venn Diagrams), Linear Arrangements
`Real-World Item`	Objects (Characters, Household Items), Contextual Scenes (Architecture, Maps, Scientific Tools)
`Cross-category`	Combinations of multiple visual categories

Table 2: Visual categories and element coverage in DeepVision-103K.

2.2 Broad Coverage

Each problem in DeepVision-103K is categorized under a hierarchical topic structure following Qiao et al. (2025). As shown in Figure 4, the dataset spans four major mathematical disciplines—**Geometry** (the largest share), **Algebra**, **Probability and Statistics**, and **Fundamental Mathematical Skills**. Beyond formula- and theorem-based mathematics, DeepVision also incorporates **visual logic problems** from Zebra-CoT (Li et al., 2025) and GameQA (Tong et al., 2025)—including visual games, maze, chess, tetris where solutions emerge primarily from visual perception and logical deduction.

3 Construction of DeepVision-103K

We curated DeepVision-103K from two open-source SFT corpora of K12-level multimodal math problems—MM-MathInstruct-3M (Wang et al., 2025b) and MultiMath-300K (Peng et al., 2024), forming an initial pool of 3.3M samples. To extract verifiable data from this large yet noisy pool, we applied a three-stage pipeline shown in Figure 5:

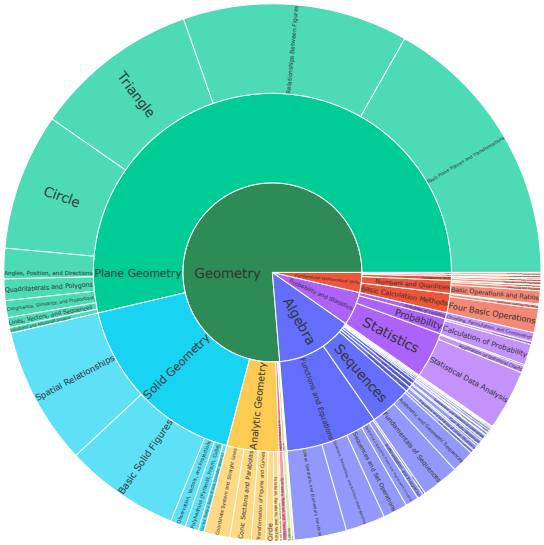


Figure 4: Mathematical topics in DeepVision-103K.

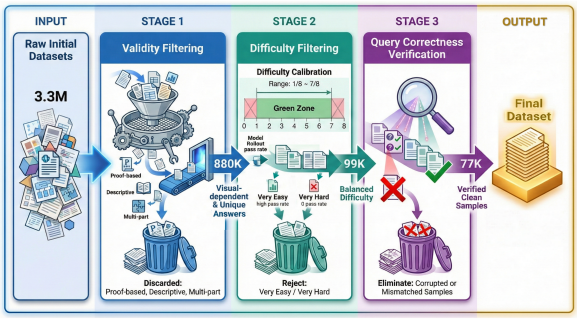


Figure 5: Math data curation pipeline.

1. **Validity Filtering:** Remove problems unsuitable for RL training, including proof-based, descriptive, or multi-answer questions.

2. **Difficulty Filtering:** Calibrate per-sample difficulty to the model via rollout pass rates.

3. **Query Correctness Verification:** Remove samples with corrupted text, image–text mismatches, or incorrect reference answers.

Stage 1: Validity Filtering. Reinforcement learning requires unique and verifiable answers to provide reward signals. In this stage, we first applied rule-based filtering to **remove proof or explanation tasks** containing keywords such as "prove", "explain", "describe". For the remaining part, we used Qwen3-VL-32B-Instruct (Bai et al., 2025) to

filter out questions with multiple answers and assess whether visual information is essential, only retained questions that satisfied both. Finally, we performed comprehensive benchmark **decontamination via exact matching and N-gram (N=32) matching** against all evaluation benchmarks. This stage yielded 880K math samples.

Stage 2: Difficulty Filtering. Data with appropriate difficulty is crucial for efficient RL training (Zeng et al., 2025b). Following Qwen3-VL (Bai et al., 2025) we derived difficulty from rollout pass rates. For each question, **we performed 8 rollouts with MiMo-VL-7B-SFT-2508** (Team et al., 2025), computed accuracy via MathVerify (Kydlíček, 2025), and kept samples whose pass rate falls in $[\frac{1}{8}, \frac{7}{8}]$. Zero-pass samples were dropped as too hard or unverifiable. Full-pass samples were also excluded for two reasons: (1) uniformly correct rollouts yield zero advantage and thus no gradient signal; (2) over-representing easy queries causes a rapid drop in policy entropy, hindering exploration of complex reasoning (Zeng et al., 2025a). However, some knowledge points appear mostly in easy samples, **we applied knowledge-guided recall to preserve their coverage** (Appendix C.1). This stage yielded 99K math samples.

Stage 3: Query Correctness Verification. Reliable RL rewards also require well-formed questions. Pass-rate filtering removes too-hard samples but cannot catch inherently problematic queries that models happen to answer correctly (e.g., with garbled text or image–text mismatches). We therefore prompted Gemini-3-Flash (Google, 2025) to (1) **verify that each question is complete and free of corrupted text** and (2) **detect image–text mismatches**; only samples passing both checks were retained (protocol in Appendix C.2). This stage yielded 77K verified mathematics samples, **an ablation in Sec. 5.3 confirms its necessity.**

Beyond mathematics, **we additionally included visual-logic data** from Zebra-CoT (Li et al., 2025) and GameQA (Tong et al., 2025). These sources are already structurally clean, we used them to broaden the data distribution rather than to scale up. We applied the same rollout filtering with a tighter window of $[\frac{3}{8}, \frac{4}{8}]$ and skipped Stage 3 verification, yielding 26K verifiable examples (details in Appendix C.1). **Combined with the 77K math samples, this produced the final DeepVision-103K.**

4 Experiments

We present a comprehensive evaluation of mathematical and general multimodal reasoning capabilities of models trained on DeepVision-103K.

4.1 Setup

Models We conducted RLVR training on Qwen3-VL-4/8B-Instruct (Bai et al., 2025) and MiMo-VL-7B-SFT-2508 (Team et al., 2025). All three backbones received visual reasoning data during pretraining or midtraining, exhibiting native visual thinking that allows direct RL training without a cold-start SFT phase. We denote the resulting models as **Q3-4B-DV**, **Q3-8B-DV**, and **MiMo-DV**.

Algorithm We employed GSPO (Zheng et al., 2025) for RL training, utilizing rule-based rewards based on answer correctness (+1 for correct answers, 0 otherwise). Response format is enforced through prompts thus no format reward was applied. Detailed training configurations and prompts are provided in Appendix D.

Baselines Three groups of baselines are considered: (1) **Closed-source models**: GPT-5-Nano-High and Gemini-2.5-Flash-Lite; (2) **Official thinking variants**: Qwen3-VL-4/8B-Thinking and MiMo-VL-7B-RL-2508; (3) **Open-source RL datasets**: MathBook (Qiao et al., 2025), a synthetic mathematical dataset; MM-Eureka (Meng et al., 2025), human-annotated K12 data; OpenMMReasoner (Zhang et al., 2025), filtration and recombination of prior sources. For group (3), we trained MiMo-VL-7B-SFT-2508 on each dataset under the same RL settings as MiMo-VL-7B-DeepVision for a fair comparison.

Evaluation Models are evaluated on (1) **Multimodal Math**: WeMath (Qiao et al., 2024), $\text{MathVerse}_{\text{vision}}$ (Zhang et al., 2024), MathVision (Wang et al., 2024), and LogicVista (Xiao et al., 2024); (2) **General Multimodal**: MMMU_{val} (Yue et al., 2024a), MMMU_{Pro} (Yue et al., 2024b), and M^3CoT (Chen et al., 2024). The maximum response length is set to 32K tokens, with decoding parameters following each model's official recommendations. Complete details are provided in Appendix E.

4.2 Mathematical Reasoning

Improvements over bases. Across all four math benchmarks in Table 3, Q3-4B-DV / Q3-8B-DV / MiMo-DV improve over their bases by +4.73 /

Model	Multimodal Math					General Multimodal			
	WeMath	MathVision	$\text{MathVerse}_{\text{vision}}$	LogicVista	Avg.	MMMU_{val}	MMMU_{Pro}	M^3CoT	Avg.
Closed-source Models									
GPT-5-Nano-High	78.62	58.75	70.30	58.03	66.43	70.78	70.64	69.15	70.19
Gemini-2.5-Flash-Lite	83.85	52.47	70.30	60.49	66.78	64.77	65.08	68.42	66.09
Qwen3-VL-4B Series									
Qwen3-VL-4B-Instruct	75.63	46.31	66.87	59.07	61.97	61.78	58.32	72.35	64.15
Qwen3-VL-4B-Thinking	81.44	**53.62**	69.25	61.52	66.46	68.33	66.53	72.43	69.10
Qwen3-VL-4B-DeepVision	**82.30**	52.11	**70.43**	**61.96**	**66.70**	**68.89**	**67.63**	**72.91**	**69.81**
Qwen3-VL-8B Series									
Qwen3-VL-8B-Instruct	79.36	51.44	67.38	61.16	64.84	67.66	67.69	70.83	68.73
Qwen3-VL-8B-Thinking	84.54	**57.89**	**72.84**	64.73	**70.00**	69.33	**70.29**	71.31	70.31
Qwen3-VL-8B-DeepVision	**85.11**	55.49	72.46	**64.73**	69.45	**71.33**	**70.29**	**71.61**	**71.08**
MiMo-VL-7B Series									
MiMo-VL-7B-SFT-2508	74.42	50.69	72.71	60.71	64.63	63.77	60.69	70.02	64.83
MiMo-VL-7B-RL-2508	76.95	53.91	**76.39**	64.28	67.88	67.44	63.87	70.57	67.29
MiMo-VL-7B-MM-Eureka	79.08	50.00	73.35	61.16	65.90	67.67	65.78	70.36	67.94
MiMo-VL-7B-MathBook	77.18	51.31	73.60	62.28	66.09	66.33	63.47	70.23	66.68
MiMo-VL-7B-OpenMMReasoner	**83.45**	52.97	74.87	61.68	68.24	66.78	66.82	**78.21**	70.60
MiMo-VL-7B-DeepVision	82.98	**55.24**	76.26	**65.62**	**70.03**	**71.00**	**69.19**	72.56	**70.92**

Table 3: Performance comparison across multimodal mathematical reasoning and general multimodal benchmarks. We report Average@8 accuracy (%). The best results within each model family are in **bold**; the second-best are underlined.

+4.61 / +5.40 on average. Per-benchmark gains span +2.89 to +8.56, with the largest jumps on WeMath (+6.67 / +5.75 / +8.56).

Matching strong baselines. On the math average, Q3-8B-DV and MiMo-DV exceed the stronger closed-source baseline (Gemini-2.5-Flash-Lite, 66.78) by +2.67 and +3.25, while Q3-4B-DV matches it (66.70 vs 66.78). Against the official thinking counterparts on the same backbones, all three remain competitive: MiMo-DV beats MiMo-VL-7B-RL by +2.15 (70.03 vs 67.88), Q3-4B-DV slightly surpasses Qwen3-VL-4B-Thinking (66.70 vs 66.46), and Q3-8B-DV stays within 0.55 points of Qwen3-VL-8B-Thinking (69.45 vs 70.00).

Superiority over open-source datasets. Under matched training (same base, same training steps), MiMo-DV wins 3 of 4 math benchmarks against MM-Eureka, MathBook, and OpenMMReasoner, tops the math average (70.03), and +1.79 over the strongest alternative OpenMMReasoner (68.24).

4.3 Generalization Beyond Mathematics

The same pattern holds on the general multimodal average. Q3-4B-DV / Q3-8B-DV / MiMo-DV lift their bases by +5.66 / +2.35 / +6.09 and beat their official thinking counterparts by +0.71 / +0.77 / +3.63; Q3-8B-DV and MiMo-DV further surpass the strongest closed-source baseline (GPT-5-Nano-High, 70.19). **MiMo-DV also tops the general average against open-source RL datasets** (70.92 vs OpenMMReasoner's 70.60, despite the latter's M^3CoT contamination). This suggests that diverse visual elements and broad domain coverage are crucial for enhancing general multimodal reasoning capabilities, which is further supported by our analysis in Sec. 5.2.

4.4 Reasoning Efficiency

Table 4 compares accuracy v.s. average token cost across Qwen3-VL-8B variants.

Two observations stand out. First, DeepVision raises the average accuracy from 66.50% to 70.15% (+3.65) over Instruct while extending response length by 36% (2,333→3,173 tokens), indicating that RLVR on DeepVision activates longer reasoning chains. Second, DeepVision matches the official Thinking variant in accuracy (70.15% vs. 70.13%) while consuming only 63.5% of the tokens (3,173 vs. 4,995). This suggests that well-curated, difficulty-calibrated training data can yield equally effective but more concise reasoning chains.

Benchmark	Instruct Acc / Tok.	DeepVision Acc / Tok.	Thinking Acc / Tok.
WeMath	79.36 / 1,428	**85.11** / 2,010	84.54 / 3,754
MathVision	51.44 / 4,288	55.49 / 5,738	**57.89** / 8,970
MathVerse	67.38 / 1,572	72.46 / 2,714	**72.84** / 4,665
LogicVista	61.16 / 1,769	**64.73** / 2,716	**64.73** / 6,115
MMMU_{val}	67.66 / 2,099	**71.33** / 2,758	69.33 / 5,082
MMMU_{Pro}	67.69 / 2,170	**70.29** / 2,895	**70.29** / 5,037
M^3CoT	70.83 / 1,029	**71.61** / 1,294	71.31 / 2,761
Average	66.50 / 2,333	**70.15** / 3,173	70.13 / 4,995

Table 4: Reasoning efficiency comparison on the Qwen3-VL-8B family. We report Average@8 accuracy (%) and the average number of response tokens.

Training Data	Multimodal Math					General Multimodal			
	WeMath	MathVision	$\text{MathVerse}_{\text{vision}}$	LogicVista	Avg.	MMMU_{val}	MMMU_{Pro}	M^3CoT	Avg.
MiMo-VL-7B-SFT-2508	74.42	50.69	72.71	60.71	64.63	63.77	60.69	70.02	64.83
DeepVision-103K_{200}	**82.98**	55.24	**76.26**	65.62	**70.03**	**71.00**	69.19	72.56	**70.92**
synergy with open-source RL datasets									
MathBook	77.18	51.31	73.60	62.28	66.09	66.33	63.47	70.23	66.68
MathBook + DeepVision-20K	81.20	53.68	74.11	63.53	68.13	68.33	67.28	72.60	69.40
MM-Eureka	79.08	50.00	73.35	61.16	65.90	67.67	65.78	70.36	67.94
MM-Eureka + DeepVision-20K	80.46	54.27	73.85	**66.21**	68.70	68.00	67.15	71.61	68.92
w/o visual logic data									
Math-77K_{150}	81.67	54.83	74.23	63.98	68.68	70.00	68.55	72.09	70.21
Math-77K_{200}	82.07	**55.72**	74.74	63.53	69.02	68.50	**69.67**	**72.65**	70.27
subsets of mathematical data									
Geometry-58K_{110}	79.59	53.55	74.11	63.98	67.81	68.77	67.51	71.22	69.17
Non-Geometry-19K_{40}	80.29	53.98	74.74	62.86	67.97	66.89	66.88	70.79	68.19
w/o multimodal math data									
Visual-logic-26K_{50}	79.54	51.61	73.35	63.98	67.12	68.33	67.34	71.61	69.09
w/o correctness verification									
Unverified-125K_{200}	82.36	53.02	73.47	62.86	67.93	69.33	67.80	71.70	69.61

Table 5: Synergy and ablation results on MiMo-VL-7B-SFT-2508 (Average@8). Bold marks the best per column across all rows.

5 Analyses

Our analyses investigate the following questions:
***Q1**: Synergy with Existing Datasets.* DeepVision-103K is curated from open-source corpora, does it provide signal that is complementary to other RL datasets, or largely overlap with them?
***Q2**: Data Composition.* How do the data domains (multimodal math, visual logic) and math sub-categories (geometry, non-geometry) each contribute to the final performance?
***Q3**: Necessity of Query Correctness Verification.* Recent studies (Wu et al., 2025; Shao et al., 2025) suggest that RLVR can work even under random rewards. Is the correctness verification step truly necessary in our data curation pipeline?
***Q4**: Enhanced Capabilities.* What capabilities are enhanced after RL on DeepVision-103K?
***Q5**: SFT effectiveness.* Is DeepVision-103K useful beyond RLVR—does it also work as SFT data?

5.1 Synergy with Existing Datasets

To test whether DeepVision contributes distinct training signal, we randomly sampled 20K examples from it and mixed them with two open-source datasets, MathBook and MM-Eureka, training MiMo-VL-7B-SFT-2508 under the same RL configuration as Section 4.

Adding 20K DeepVision samples lifts MathBook (+2.04 / +2.72) and MM-Eureka (+2.80 / +0.98) on math / general average, with consistent per-benchmark gains. **This indicates that DeepVision contributes signal the existing datasets do not already cover.**

5.2 Data Composition

Math vs. Visual Logic. DeepVision spans two data domains with different reasoning paradigms: multimodal math requires extracting visual evidence and applying mathematical knowledge (formulas, theorems, computations), whereas visual logic relies mainly on visual cues such as object positions, spatial relations, and patterns. As Zha et al. (2025) notes that mixing heterogeneous domains may introduce conflicting gradients, we examine whether the visual-logic data is indeed beneficial and how each domain contributes.

To isolate each domain's contribution, we compared the full mixed setting (DeepVision-103K_{200}) against three single-domain ablations:

- **Math-77K_{150}**: math only for 150 steps (same math exposure as DeepVision_{200}).
- **Math-77K_{200}**: math only for 200 steps (same total exposure as DeepVision_{200}).
- **Visual-logic-26K_{50}**: visual logic only for 50 steps (same visual-logic exposure as DeepVision_{200}).

Table 5 shows that **scaling math training is consistently beneficial**: both math-only variants outperform the base, and extending from 150 to 200 steps improves every benchmark. However, **math alone is not sufficient**: Math-77K_{200} underperforms the mixed setting on the math average (69.02 vs 70.03), with a clear gap on LogicVista (63.53 vs 65.62). **Visual-logic alone also improves** over the base on every benchmark, demonstrating

positive transfer from visual logic to both mathematical and general evaluations. Crucially, **two domains exhibit positive synergy**: the mixed setting strictly outperforms both single-domain ablations on math (70.03 vs 69.02 / 67.12) and general (70.92 vs 70.27 / 69.09) averages. We attribute these gains to two factors: (1) spatial reasoning and pattern recognition are broadly useful primitives shared across mathematical and general multimodal tasks; (2) visual-logic training directly strengthens these primitives, while multimodal math alone does not sufficiently cultivate them.

Geometry vs. Non-Geometry. Geometry constitutes the majority of our mathematical subset, while the remaining non-geometry categories (algebra, statistics, fundamental skills) account for only 19K samples, raising the question of whether they are dispensable. We split Math-77K into Geometry-58K and Non-Geometry-19K and trained each independently with steps proportional to its data volume. Results shows that **both subsets help individually**: each improves over the base on math and general averages. **The much smaller non-geometry subset is effective**: Non-Geometry-19K alone delivers +3.34 / +3.36 on math / general averages, on par with the $\sim 3\times$ larger Geometry-58K. **Their combination outperforms either subset alone**: the union (Math-77K$_{150}$) reaches 68.68 / 70.21 on math / general averages, exceeding both Geometry-58K (67.81 / 69.17) and Non-Geometry-19K (67.97 / 68.19). This suggests the two subsets are complementary rather than redundant.

5.3 Necessity of Correctness Verification

After difficulty filtering, we obtained 99K samples calibrated to model capability. Stage 3 verification further removed an additional 22K samples with garbled text or image–text mismatches. The ablation in Table 5 (*Unverified-125K*$_{200}$) shows that **skipping this stage hurts performance**: math drops 70.03→67.93 (−2.10) and general 70.92→69.61 (−1.31), confirming that query verification is essential for clean training data.

5.4 Enhanced Capabilities

Training on DeepVision-103K presents increasing response length, upward rewards, and stable entropy (Appendix F). To investigate *how* DeepVision improves over its base, we collected cases where Qwen3-VL-8B-DeepVision succeeds but Qwen3-VL-8B-Instruct fails, and asked human annotators to label the improvement mechanism following a structured protocol (Appendix H). Our analysis reveals three enhancement types (Figure 6).

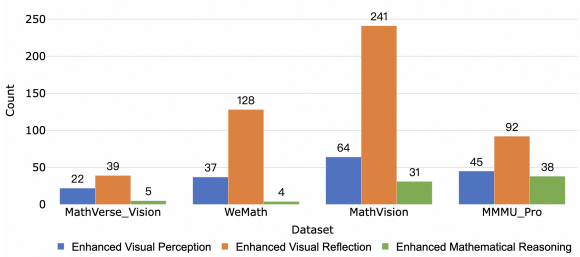


Figure 6: Enhanced Capabilities

Type I: Enhanced Visual Perception. DeepVision model exhibits **stronger "one-shot perception"**—it correctly identifies geometric features on the initial observation (Figure 7).

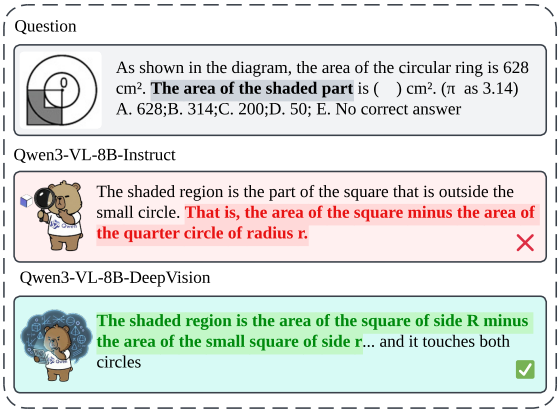


Figure 7: DeepVision correctly identifies the shaded region on the first attempt.

Type II: Enhanced Visual Reflection. When initial perception falters, DeepVision model **conducts genuine visual re-examination**, whereas the Instruct model rephrases its conclusions (Figure 8).

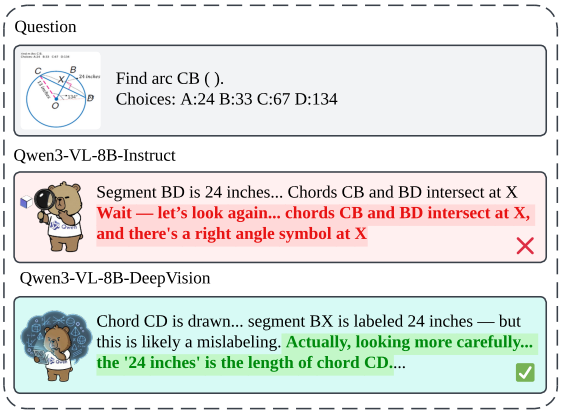


Figure 8: DeepVision actively re-examines visual content to correct errors, while the Instruct model merely rephrases without genuine verification.

Type III: Enhanced Mathematical Reasoning. Beyond visual capabilities, RL training enhances mathematical reasoning. In cases where both models correctly extract identical visual information, DeepVision **demonstrates more rigorous deduction** (Figure 9).

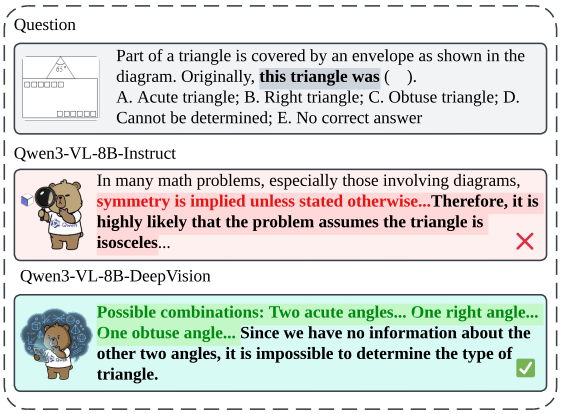


Figure 9: DeepVision systematically enumerates all possible angle combinations and concludes the type cannot be determined.

We note that DeepVision also introduces a small set of drawbacks over the Instruct baseline. Two patterns recur: (i) *Overthinking*: on easy questions the Instruct model solves in a few lines, DeepVision sometimes reaches for coordinates, case splits, or formal arguments, and talks itself out of the right answer. (ii) *Visual drift in extended reasoning*: as the chain of thought grows longer, the model's grip on the image can slip—it misreads a coordinate it had earlier read correctly, miscounts after a long detour. We view both as the flip side of the same exploratory behaviors that drive the gains above.

5.5 Effectiveness as SFT Data

We distilled CoTs from Qwen3-VL-235B-A22B-Instruct on DeepVision-103K queries and conducted SFT on Qwen2.5-VL-7B-Instruct. For each query we selected up to three correct rollouts under three strategies: **longest** (3 longest), **shortest** (3 shortest), and **diverse** (3 spanning the widest length range); queries with fewer than three correct rollouts retain all of them. Each strategy yields 227K SFT pairs (pass-rate distribution in Appendix G, also provide passrate refe for RL training on larger size backbones). As a controlled baseline, we include **OpenMMReasoner-ColdStart** (+OMR-CS) (Zhang et al., 2025), which uses the same teacher–student pair but distills on OpenMMReasoner queries and attains 874K SFT pairs, isolating the effect of query quality.

Benchmark	Instruct	+OMR-CS	+DV-SFT		
			longest	diverse	shortest
Multimodal Math					
WeMath	60.85	71.23	**78.28**	78.02	73.88
MathVision	16.94	**34.60**	34.14	33.35	32.40
$\text{MathVerse}_{\text{vision}}$	41.62	50.92	53.27	**54.98**	52.66
LogicVista	43.29	48.77	**51.55**	50.95	48.21
Avg.	40.67	51.38	54.31	**54.33**	51.79
Δ	–	+10.71	+13.64	**+13.65**	+11.11
General Multimodal					
MMMU_{val}	51.59	52.59	**57.63**	56.48	54.91
MMMU_{Pro}	49.73	51.59	54.64	**55.40**	54.81
M^3CoT	64.42	**67.95**	67.02	66.50	66.43
Avg.	55.25	57.38	**59.76**	59.46	58.72
Δ	–	+2.13	**+4.52**	+4.21	+3.47

Table 6: SFT cold-start on Qwen2.5-VL-7B-Instruct (Avg@8). +OMR-CS and +DV-SFT share the same teacher (Qwen3-VL-235B-A22B-Instruct); best in **bold**.

We evaluated all SFT checkpoints under identical inference settings (temperature 0.5, top-p 0.95). All three DV-SFT variants consistently outperform OMR-CS on average despite using only 227K vs. 874K SFT pairs (Table 6), with the best variant leading by +1.93 on math and +2.39 on general benchmarks. This confirms DeepVision-103K's SFT effectiveness. Among strategies, **longest** leads or ties on most benchmarks, indicating that longer correct CoTs provide richer supervision for visual reasoning cold-start.

6 Conclusion

We present **DeepVision-103K**, a visually diverse and broad-coverage multimodal dataset for RLVR, curated from real-world K12 sources via validity filtering, pass-rate difficulty calibration, and correctness verification. Models trained on DeepVision-103K achieve up to +5.39% math and +6.09% general gains, matching official thinking variants while using only 63.5% of the tokens. Controlled ablations confirm that multimodal math and visual logic exhibit positive synergy, correctness verification is indispensable, and DeepVision provides complementary signal to existing datasets. Human annotation further identifies enhanced visual perception, visual reflection, and mathematical reasoning as the key improvements. Sec 5.5 demonstrates DeepVision-103K's effectiveness as SFT data. We release DeepVision-103K to facilitate future research in multimodal reasoning.

7 Limitations

DeepVision-103K focuses on K12-level problems with unique verifiable answers, so it does not cover open-ended tasks such as proof writing or multi-solution problems. The visual distribution, while substantially broader than prior datasets, remains imbalanced—planar geometry dominates, and rare element types (e.g., schematic diagrams) are under-represented. The difficulty calibration is performed with MiMo-VL-7B-SFT-2508 and validated at 4B–8B scales; extending to 32B+ backbones is left to future work due to compute constraints. Finally, the correctness verification stage relies on Gemini-3-Flash, which may introduce mild model-specific bias despite its clear net benefit.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Qiguang Chen, Libo Qin, Jin Zhang, Zhi Chen, Xiao Xu, and Wanxiang Che. 2024. M^3cot: A novel benchmark for multi-domain multi-step multi-modal chain-of-thought. *Preprint*, arXiv:2405.16473.

Yew Ken Chia, Vernon Toh Yan Han, Deepanway Ghosal, Lidong Bing, and Soujanya Poria. 2024. Puzzlevqa: Diagnosing multimodal reasoning challenges of language models with abstract visual patterns. *Preprint*, arXiv:2403.13315.

Google. 2025. Gemini3-flash-preiview model card. https://deepmind.google/models/gemini/flash/.

Hynek Kydlíček. 2025. Math-Verify: Math Verification Library.

Ang Li, Charles Wang, Kaiyu Yue, Zikui Cai, Ollie Liu, Deqing Fu, Peng Guo, Wang Bill Zhu, Vatsal Sharan, Robin Jia, Willie Neiswanger, Furong Huang, Tom Goldstein, and Micah Goldblum. 2025. Zebra-cot: A dataset for interleaved vision language reasoning. *Preprint*, arXiv:2507.16746.

Wentao Liu, Qianjun Pan, Yi Zhang, Zhuo Liu, Ji Wu, Jie Zhou, Aimin Zhou, Qin Chen, Bo Jiang, and Liang He. 2024. Cmm-math: A chinese multimodal math dataset to evaluate and enhance the mathematics reasoning of large multimodal models. *Preprint*, arXiv:2409.02834.

Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. Inter-gps: Interpretable geometry problem solving with formal language and symbolic reasoning. *Preprint*, arXiv:2105.04165.

Fanqing Meng, Lingxiao Du, Zongkai Liu, Zhixiang Zhou, Quanfeng Lu, Daocheng Fu, Tiancheng Han, Botian Shi, Wenhai Wang, Junjun He, Kaipeng Zhang, Ping Luo, Yu Qiao, Qiaosheng Zhang, and Wenqi Shao. 2025. Mm-eureka: Exploring the frontiers of multimodal reasoning with rule-based reinforcement learning. *Preprint*, arXiv:2503.07365.

Kaichun Mo, Shilin Zhu, Angel X. Chang, Li Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. 2018. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. *Preprint*, arXiv:1812.02713.

Shuai Peng, Di Fu, Liangcai Gao, Xiuqin Zhong, Hongguang Fu, and Zhi Tang. 2024. Multimath: Bridging visual and mathematical reasoning for large language models. *Preprint*, arXiv:2409.00147.

Yingzhe Peng, Gongrui Zhang, Miaosen Zhang, Zhiyuan You, Jie Liu, Qipeng Zhu, Kai Yang, Xingzhong Xu, Xin Geng, and Xu Yang. 2025. Lmm-r1: Empowering 3b lmms with strong reasoning abilities through two-stage rule-based rl. *Preprint*, arXiv:2503.07536.

Runqi Qiao, Qiuna Tan, Guanting Dong, Minhui Wu, Chong Sun, Xiaoshuai Song, Zhuoma GongQue, Shanglin Lei, Zhe Wei, Miaoxuan Zhang, Runfeng Qiao, Yifan Zhang, Xiao Zong, Yida Xu, Muxi Diao, Zhimin Bao, Chen Li, and Honggang Zhang. 2024. We-math: Does your large multimodal model achieve human-like mathematical reasoning? *Preprint*, arXiv:2407.01284.

Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. 2025. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. *Preprint*, arXiv:2508.10433.

PAUL L. Rosin. 2008. *2D Shape Measures for Computer Vision*, pages 347–371.

Rulin Shao, Shuyue Stella Li, Rui Xin, Scott Geng, Yiping Wang, Sewoong Oh, Simon Shaolei Du, Nathan Lambert, Sewon Min, Ranjay Krishna, Yulia Tsvetkov, Hannaneh Hajishirzi, Pang Wei Koh, and Luke Zettlemoyer. 2025. Spurious rewards: Rethinking training signals in rlvr. *Preprint*, arXiv:2506.10947.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, and 55 others. 2025. Mimo-vl technical report. *Preprint*, arXiv:2506.03569.

Jingqi Tong, Jixin Tang, Hangcheng Li, Yurong Mou, Ming Zhang, Jun Zhao, Yanbo Wen, Fan Song, Jiahao Zhan, Yuyang Lu, Chaoran Tao, Zhiyuan Guo, Jizhou Yu, Tianhao Cheng, Zhiheng Xi, Changhao Jiang, Zhangyue Yin, Yining Zheng, Weifeng Ge, and 5 others. 2025. Game-rl: Synthesizing multimodal verifiable game data to boost vlms' general reasoning. *Preprint*, arXiv:2505.13886.

Haozhe Wang, Chao Qu, Zuming Huang, Wei Chu, Fangzhen Lin, and Wenhu Chen. 2025a. Vl-rethinker: Incentivizing self-reflection of vision-language models with reinforcement learning. *Preprint*, arXiv:2504.08837.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *Preprint*, arXiv:2402.14804.

Ke Wang, Junting Pan, Linda Wei, Aojun Zhou, Weikang Shi, Zimu Lu, Han Xiao, Yunqiao Yang, Houxing Ren, Mingjie Zhan, and Hongsheng Li. 2025b. Mathcoder-VL: Bridging vision and code for enhanced multimodal mathematical reasoning. In *The 63rd Annual Meeting of the Association for Computational Linguistics*.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Hongjin Lu, Linjie Li, Chung-Ching Lin, Kevin Lin, Furong Huang, and Lijuan Wang. 2025c. Sota with less: Mcts-guided sample selection for data-efficient visual reasoning self-improvement. *Preprint*, arXiv:2504.07934.

Mingqi Wu, Zhihao Zhang, Qiaole Dong, Zhiheng Xi, Jun Zhao, Senjie Jin, Xiaoran Fan, Yuhao Zhou, Huijie Lv, Ming Zhang, Yanwei Fu, Qin Liu, Songyang Zhang, and Qi Zhang. 2025. Reasoning or memorization? unreliable results of reinforcement learning due to data contamination. *Preprint*, arXiv:2507.10532.

Jiaer Xia, Yuhang Zang, Peng Gao, Sharon Li, and Kaiyang Zhou. 2025. Visionary-r1: Mitigating shortcuts in visual reasoning with reinforcement learning. *Preprint*, arXiv:2505.14677.

Yijia Xiao, Edward Sun, Tianyu Liu, and Wei Wang. 2024. Logicvista: Multimodal llm logical reasoning benchmark in visual contexts. *Preprint*, arXiv:2407.04973.

Senqiao Yang, Junyi Li, Xin Lai, Bei Yu, Hengshuang Zhao, and Jiaya Jia. 2025a. Visionthink: Smart and efficient vision language model via reinforcement learning. *Preprint*, arXiv:2507.13348.

Yi Yang, Xiaoxuan He, Hongkun Pan, Xiyan Jiang, Yan Deng, Xingtao Yang, Haoyu Lu, Dacheng Yin, Fengyun Rao, Minfeng Zhu, Bo Zhang, and Wei Chen. 2025b. R1-onevision: Advancing generalized multimodal reasoning through cross-modal formalization. *Preprint*, arXiv:2503.10615.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, and 3 others. 2024a. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *Preprint*, arXiv:2311.16502.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. 2024b. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. *arXiv preprint arXiv:2409.02813*.

Weihao Zeng, Yuzhen Huang, Qian Liu, Wei Liu, Keqing He, Zejun Ma, and Junxian He. 2025a. Simplerl-zoo: Investigating and taming zero reinforcement learning for open base models in the wild. *Preprint*, arXiv:2503.18892.

Yongcheng Zeng, Zexu Sun, Bokai Ji, Erxue Min, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Haifeng Zhang, Xu Chen, and Jun Wang. 2025b. Cures: From gradient analysis to efficient curriculum learning for reasoning llms. *Preprint*, arXiv:2510.01037.

Yuheng Zha, Kun Zhou, Yujia Wu, Yushu Wang, Jie Feng, Zhi Xu, Shibo Hao, Zhengzhong Liu, Eric P. Xing, and Zhiting Hu. 2025. Vision-g1: Towards general vision language reasoning with multi-domain data curation. *Preprint*, arXiv:2508.12680.

Kaichen Zhang, Keming Wu, Zuhao Yang, Bo Li, Kairui Hu, Bin Wang, Ziwei Liu, Xingxuan Li, and Lidong Bing. 2025. Openmmreasoner: Pushing the frontiers for multimodal reasoning with an open and general recipe. *Preprint*, arXiv:2511.16334.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, and Hongsheng Li. 2024. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *Preprint*, arXiv:2403.14624.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

A Visual Examples

In this section, we present cross-category visual combination examples in DeepVision-103K.

B Visual Elements Annotation

To characterize the distribution of visual elements in DeepVision-103K and existing datasets, we constructed a visual annotation taxonomy based on (Mo et al., 2018; Rosin, 2008). We then instructed GPT-5-mini to annotate visual elements in each dataset according to the proposed taxonomy

Question: Water is poured into a cup (the rate of water pouring remains constant), the relationship between the height of water h in the cup and the pouring time t is shown in the figure. What could be the shape of the cup?
A) ①③ B) ①② C) ①②③ D) ②③

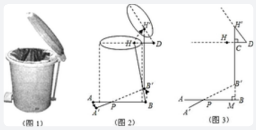
Image：

Visual Elements:
Solid Geometry: Hemisphere, Cone, Cylinder, Frustum
Analytic Plots: Coordinate System, Function Curve

Figure 10: Solid Geometry & Analytic Plots.

Question: A household trash can has a horizontal pedal \(AB\) that rotates about a fixed point \(P\)....In the opened position (Figure 3), t find the minimum height of the pedal \(AB\) above the ground so that \(\angle HDH' \ge 60^\circ\). (Answer to two significant figures; \(\sqrt2\approx1.41,\ \sqrt3\approx1.73\).)

Image：



Visual Elements:
Planar Geometry: Triangle, Angles, Parallel Lines, Perpendicular Lines
Solid Geometry: Cylinder
Real-World Item: Household Item

Figure 11: Planar Geometry & Solid Geometry & Real-World Item.

(Table 7). We set the decoding temperature to 0.1 and the thinking budget to `low`.

C Data Construction

C.1 Difficulty Filtering

For the math subset, in addition to the kept window $[\frac{1}{8}, \frac{7}{8}]$ defined in Stage 2, we selectively recalled full-pass samples (8/8) for knowledge points that were under-represented in the kept window, preserving coverage while keeping the dataset size manageable.

Knowledge-guided recall. For every fine-grained knowledge point k, we computed its frequency in the full-pass region (pass rate $= 1$) and in the kept window ($[\frac{1}{8}, \frac{7}{8}]$), and defined a *gap ratio*

$$g(k) = \frac{f_{\text{full}}(k) - f_{\text{kept}}(k)}{f_{\text{full}}(k) + \epsilon},$$

which measures how under-represented a knowledge point is in the kept window relative to the full-pass pool. We triggered recall for every knowledge point with $g(k) > 0.35$. The threshold was chosen empirically: large enough to ignore mild long-tail fluctuations, while still capturing heavily under-represented domains (e.g., conic sections and tangency reasoning).

To isolate the contribution of knowledge-guided recall, we trained MiMo-VL-7B-SFT-2508 on the Stage 2 math subset (math-only, no Stage 3 verification) with and without the 9.8K recalled samples, under matched training steps. Table 8 shows that removing the recalled samples drops the math average by 0.90 (68.48→67.58) with consistent per-benchmark degradation, confirming that recall provides a reliable gain by covering otherwise under-represented knowledge points. The top 10 recalled knowledge points are shown in Figure 12 and Table 9.

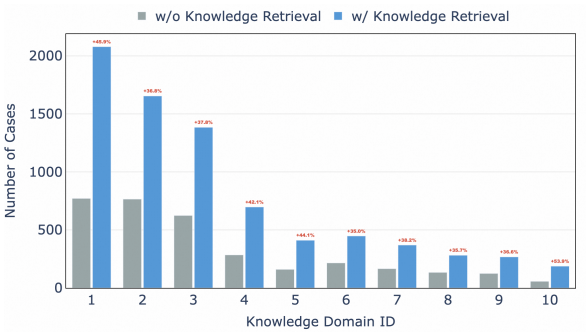


Figure 12: Top 10 Knowledge-based retrieval. The x-axis IDs correspond to knowledge domains listed in Table 9.

For visual logic data, we used tetris, maze, and chess data from Zebra-CoT (Li et al., 2025) and game data from GameQA (Tong et al., 2025) with pass rate in $[\frac{3}{8}, \frac{4}{8}]$, broadening the training data distribution while keeping the dataset size manageable.

Deployment to other model scales. For backbones with capabilities significantly different from the calibration model (e.g., 32B+), we recommend two complementary strategies that do not require re-

Category	Fine-grained types
planar_geometry	Right Triangle; Equilateral Triangle; Triangle; Square; Rectangle; Rhombus; Parallelogram; Trapezoid; Quadrilateral; Circle; Semicircle; Sector; Arc; Parallel Lines; Perpendicular Lines; Tangent; Chord; Angle; Right Angle
solid_geometry	Cube; Cuboid; Prism; Pyramid; Tetrahedron; Sphere; Cylinder; Cone; Frustum; Hemisphere; Net; Orthographic View
analytic_plot	Linear Graph; Parabola; Hyperbola; Sinusoidal Curve; Exponential Curve; Analytic Circle; General Function Curve; Coordinate System; Number Line; Scatter Points; Inequality Region; Equation
data_chart	Table; Bar Chart; Line Chart; Pie Chart; Donut Chart; Histogram; Box Plot; Stem-and-Leaf Plot
schematic_diagram	Flowchart; Circuit; Force Diagram; Tree Diagram; Venn Diagram; Linear Arrangement
real-world item	Character; Plant; Scientific Tool; Vehicle; Architecture; Household Item; Apparel; Food; Real Object; Scene; Map

Table 7: Visual-element annotation taxonomy used in this work.

Benchmark	MiMo	Stage 2 math	
		w/o recall (89K)	w/ recall (99K)
WeMath	74.42	80.23	**81.15**
MathVision	50.69	52.88	**53.42**
$\text{MathVerse}_{\text{vision}}$	72.71	73.61	**74.38**
LogicVista	60.71	63.61	**64.96**
Avg.	64.63	67.58	**68.48**
Δ	–	+2.95	**+3.85**

Table 8: Knowledge-guided recall ablation on the Stage 2 math subset (Average@8 %). Adding the 9.8K recalled samples lifts the math average by 0.90.

curating the entire dataset. (i) *Static recalibration*: re-run rollouts on the raw mathematical queries with the target model and re-apply Stage 2 to obtain a backbone-specific pass-rate distribution. (ii) *Dynamic group filtering*: enable group-level filtering during the RL rollout phase (used in our setting, see Appendix D), which discards groups whose rollouts are entirely correct or entirely incorrect, ensuring that gradient updates are computed only on samples with informative reward variance.

C.2 Correctness Verification

To ensure data reliability, we used GEMINI 3 FLASH as an automated verifier. For each instance, it jointly inspected the input image and question text and outputs a *label* with a *judge trace* (Table 10). The verifier rejects the instance if any of the following labels is triggered: ERR_IMG_MISSING, ERR_TEXT_MISSING, or ERR_MISMATCH. An instance is retained only when no error is detected.

C.3 Data Licenses

We list the data collection protocol of our data sources in Table 11.

D Training Details

We used verl as the training framework. Configurations for training DeepVision series models are listed in Table 12.

We used 32 H20 GPU for a single training, a training step cost 0.5h. We used the following prompt template during training and evaluation.

Training / Evaluation Prompt Template

You are a multimodal reasoning assistant. You receive images and texts, perform step-by-step reasoning (including re-checking the image) before producing the final answer. Please provide a clear, concise answer inside \boxed{} tag. For multiple choice questions, put only the letter like \boxed{A} without any additional text. For fill-in-the-blank and problem-solving questions, put only the final answer.

E Evaluation Details

We provide detailed information about the benchmarks used for evaluation and the inference hyperparameters for each model.

E.1 Benchmarks

We evaluated our models across three categories of benchmarks, as summarized in Table 13.

E.2 Inference Hyperparameters

We used different inference hyperparameters for different model families to ensure optimal performance. The detailed configurations are listed in Table 14.

For Qwen3-VL-DeepVision models, we adopted the same hyperparameters as Qwen3-VL-Instruct. For MiMo-VL-DeepVision, we adopted the same hyperparameters as MiMo-VL.

E.3 Evaluation Method

For each benchmark, we first calculated accuracy with MathVerify (Kydlíček, 2025), then prompted GPT-5-mini to re-judge cases marked as incorrect

ID	Knowledge Domain	w/o Retrieval	w/ Retrieval	Increase
1	Circle → Inscribed and Circumscribed	771	2,079	+1,308
2	Triangle → Angle of Elevation and Depression	765	1,654	+889
3	Circle → Tangency	624	1,384	+760
4	Circle → Perpendicular Chord Theorem	284	697	+413
5	Conic Sections → Hyperbola	159	410	+251
6	Figure Relationships → Inscribed and Circumscribed	215	447	+232
7	Spatial Relationships → Parallelism & Perpendicularity	165	369	+204
8	Conic Sections → Parabola	133	281	+148
9	Triangle → Criteria for Similar Triangles	124	267	+143
10	Spatial Relationships → Angle between Line & Plane	56	187	+131

Table 9: Top 10 Knowledge Domains by Retrieval Gap

Label	Category	Trigger
`ERR_IMG_MISSING`	Image quality issue	Image is missing, unreadable, or lacks essential visual information.
`ERR_TEXT_MISSING`	Missing text	Question text misses key conditions/values, making the task unsolvable.
`ERR_MISMATCH`	Image–text mismatch	Image content conflicts with the question statement.
`1`	Pass	Image and text are complete and consistent.

Table 10: Verification labels used by GEMINI 3 FLASH. Exactly one label is returned per instance.

Data Source	License	URL
MM-MathInstruct-3M (Wang et al., 2025b)	Apache 2.0	https://huggingface.co/datasets/MathLLMs/MM-MathInstruct
MultiMath-300K (Peng et al., 2024)	Unset	https://huggingface.co/datasets/pengshuai-rin/multimath-300k
Zebra-CoT (Li et al., 2025)	CC BY-NC 4.0	https://huggingface.co/datasets/multimodal-reasoning-lab/Zebra-CoT
GameQA-140K(Tong et al., 2025)	MIT	https://huggingface.co/datasets/Code2Logic/GameQA-140K
PuzzleVQA(Chia et al., 2024)	Unset	https://huggingface.co/datasets/declare-lab/PuzzleVQA

Table 11: Licenses and usage permissions for the data sources used in this work.

by MathVerify to reduce false negatives caused by parsing errors, equivalent expressions, or formatting variations. We used the revised judgment as the final label.

We deliberately chose GPT-5-mini rather than Gemini-3-Flash for evaluation re-judging. Since Gemini-3-Flash had already been used in Stage 3 of our data curation pipeline (Section 3), reusing it as the evaluation judge would risk self-confirmation bias, where the judge favors patterns it had previously selected. Using a different model family (GPT-5) provides a more independent assessment, while GPT-5-mini also offers sufficient verification capability with lower latency and cost than full-size verifiers.

F Training Curves

This section presents the training dynamics on DeepVision-103K, including response length (Figure 13), trainset rewards (Figure 14) and entropy (Figure 15).

G Distillation Teacher Pass Rate

Figure 16 shows the pass-rate distribution of Qwen3-VL-235B-A22B-Instruct on DeepVision-103K queries, used in Sec. 5.5 for SFT distillation.

H Human Annotation Protocol

Algorithm 1 details the structured rubric used in Sec. 5.4 for diagnosing why DeepVision succeeds when the Instruct base fails. For each pair of

Config	Value
lr	1e-6
kl_coef	1e-3
max_prompt_length	2K
max_response_length	16K
gen_batch_size	512
train_batch_size	256
mini_batch_size	64
micro_batch_size	32
group_filtering	acc
clip_ratio_low	1e-4
clip_ratio_high	1e-3
temperature	1.0
rollout.n	16
total_training_steps	200

Table 12: Configurations for training DeepVision series models.

Category	Benchmark	#Samples	Reference
Multimodal Math	WeMath	1,740	(Qiao et al., 2024)
	MathVision	3,040	(Wang et al., 2024)
	$\text{MathVerse}_{\text{vision}}$	788	(Zhang et al., 2024)
	LogicVista	448	(Xiao et al., 2024)
General Multimodal	M^3CoT	2,318	(Chen et al., 2024)
	MMMU_{val}	900	(Yue et al., 2024a)
	$\text{MMMU}_{\text{Pro_full}}$	1,730	(Yue et al., 2024b)

Table 13: Overview of evaluation benchmarks.

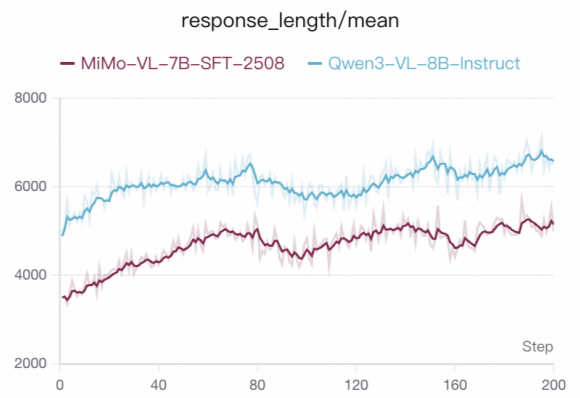


Figure 13: Increasing response length.

Parameter	Qwen3-VL-Thinking	Qwen3-VL-Instruct	MiMo-VL-(SFT/RL)
top_p	0.95	0.8	0.95
top_k	20	20	–
temperature	1.0	0.7	0.3
repetition_penalty	1.0	1.0	–
presence_penalty	0.0	1.5	–
max_tokens	32,768	32,768	32,768

Table 14: Inference hyperparameters for each model family.

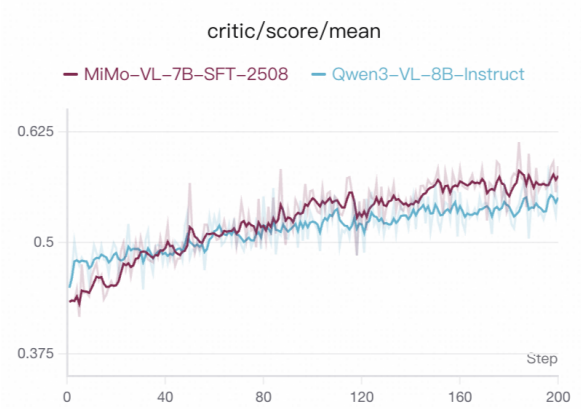


Figure 14: Upward rewards.

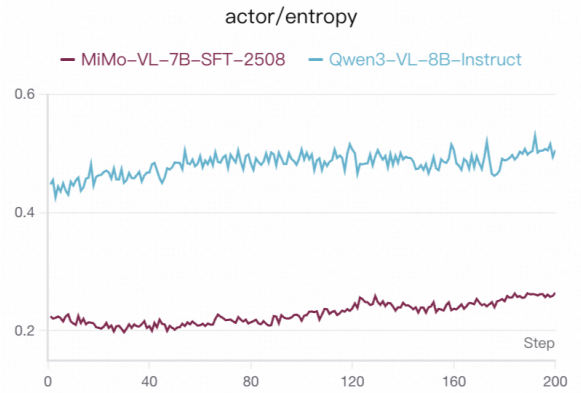


Figure 15: Stable entropy.

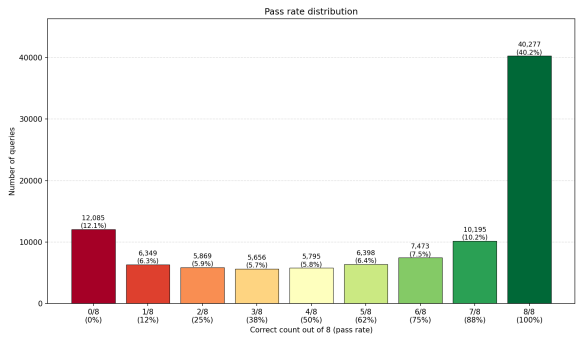


Figure 16: Pass-rate distribution of Qwen3-VL-235B-Instruct on DeepVision-103K queries (8 rollouts per query).

responses, annotators first determine whether Instruct's failure stems from visual misperception or incorrect reasoning, then identify the mechanism by which DeepVision succeeds.

I Potential Risks

We do not anticipate significant potential risks from this work. DeepVision-103K is derived from publicly available K12-level educational content and is designed for verifiable-answer multimodal reasoning rather than sensitive decision-making. The

Algorithm 1: Human Annotation Protocol

Input: Query (Image, Text), Ground Truth y, Incorrect Instruct Response R_I, Correct DeepVision Response R_D

Output: Improvement Mechanism C

```
Analyze visual descriptions in R_I
if Descriptions contradict Image then
    Root Cause ← Visual Misperception
else
    Root Cause ← Incorrect Reasoning
end
if Root Cause is Visual Misperception then
    if R_D correct at first observation then
        C ← VISUAL PERCEPTION
    else
        if R_D corrected via reflection then
            C ← VISUAL REFLECTION
        else
            C ← GUESS
        end
    end
else if Root Cause is Incorrect Reasoning then
    if R_D shows valid reasoning chain then
        C ← REASONING
    else
        C ← GUESS
    end
end
return C
```

dataset contains no personal identifiers, and our curation process filters out corrupted or unsafe samples.